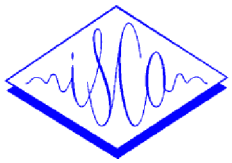

A Bayesian Model of Bilingual Segmentation for Transliteration

Andrew Finch and Eiichiro Sumita

Language Translation Group, Knowledge Creating Communication Research Center,
NICT, 3-5 Hikaridai, Keihanna Science City, Kyoto, 619-0289, Japan
{andrew.finch,eiichiro.sumita}@nict.go.jp

Abstract

In this paper we propose a novel Bayesian model for unsupervised bilingual character sequence segmentation of corpora for transliteration. The system is based on a Dirichlet process model trained using Bayesian inference through blocked Gibbs sampling implemented using an efficient forward filtering/backward sampling dynamic programming algorithm. The Bayesian approach is able to overcome the overfitting problem inherent in maximum likelihood training. We demonstrate the effectiveness of our Bayesian segmentation by using it to build a translation model for a phrase-based statistical machine translation (SMT) system trained to perform transliteration by monotonic transduction from character sequence to character sequence. The Bayesian segmentation was used to construct a phrase-table and we compared the quality of this phrase-table to one generated in the usual manner by the state-of-the-art GIZA++ word alignment process used in combination with phrase extraction heuristics from the MOSES statistical machine translation system, by using both to perform transliteration generation within an identical framework. In our experiments on English-Japanese data from the NEWS2010 transliteration generation shared task, we used our technique to bilingually co-segment the training corpus. We then derived a phrase-table from the segmentation from the sample at the final iteration of the training procedure, and the resulting phrase-table was used to directly substitute for the phrase-table extracted by using GIZA++/MOSES. The phrase-table resulting from our Bayesian segmentation model was approximately 30% smaller than that produced by the SMT system's training procedure, and gave an increase in transliteration quality measured in terms of both word accuracy and F-score.

1. Introduction

It is possible to couch the problem of transliteration as a problem of machine translation at the character level. In this paradigm, decoding is usually assumed to proceed in a monotone order, but otherwise the technique remains essentially the same, except that the tokens used in the system are characters rather than words. Recently systems based on phrase-based statistical machine translation technology are being actively researched [1, 2, 3], and have achieved state-of-the-art performance on this task. The approach makes no linguistic assumptions about the data and no intermediate phonetic representation is required, because the transduction is directly from grapheme to grapheme.

At the core of all phrase-based statistical machine translation systems (SMT) is the phrase-table. This table is the basic set of building blocks that are used to construct the translation. The creation of a phrase-table during a typical training procedure for a phrase-based SMT system consists of the following steps:

1. Word alignment using GIZA++ [4]
2. Phrase-pair extraction using heuristics (for example $grow\text{-}diag\text{-}final\text{-}and$ from the MOSES [5] toolkit)

This approach works very well in practice, but a more elegant solution would be to arrive at a set of bilingual sequence-pairs (we use this term to describe analogue of the phrase-pair at the character level) in one step, from a generative model. Unfortunately, when traditional methods that use the EM algorithm to maximize likelihood are applied to the task, they produce solutions that can grossly over-fit the data. As an extreme example, the most likely segmentation of a corpus into sequence-pairs, assuming no limits on sequence-pair size would be the entire corpus as a single bilingual sequence-pair, holding all the probability mass.

GIZA++ mitigates this problem by aligning the words in a one-to-many fashion. The single word on one side of the alignment acts as a constraint on the size of the bilingual pairs. A similar approach can be taken in transliteration, where a single character in one language is permitted to align to multiple characters of the other, but not vice versa. This approach is reasonable for English-Chinese transliteration [6, 7], where one Chinese character can be assumed to map to several English characters.

In GIZA++ this one-to-many alignment is done twice: from both source-to-target and also from target-to-source. A table of word-to-word alignments is then constructed from (typically the intersection) both of these alignments. Additional word alignments that are not in the intersection are added based on evidence and heuristics, and finally all possible phrase-pairs are extracted from the table of alignments that are consistent with the table.

In [8, 9] many-to-many alignment is performed directly using maximum likelihood training, but evidence trimming

heuristics that exclude part of the available training data are required to prevent the models from overfitting the data. [10] have successfully applied a similar Bayesian technique to grammar induction. [11] tackle the overfitting problem in phrasal alignment by using a leave-one-out approach using a strategy that despite being a different paradigm, shares many of the characteristics of our approach.

In this paper we extend existing monolingual word segmentation models ([12, 13]) to bilingual segmentation, and provide a simple yet elegant way to directly segment a bilingual training corpus in a many-to-many fashion without overfitting, using a Bayesian model.

This paper is organized as follows. In Section 2 we describe the Bayesian model used in our transliteration system. Here we give an overview of the Dirichlet process model, the Chinese Restaurant process and explain how our model relates to these two representations. We also describe the blocked Gibbs sampling technique used to train the model. In Section 4 we describe the experiments we performed to evaluate our model: the data sets, the baseline system and the training procedure. Section 5 contains the experimental results, and in Section 6 we conclude and mention promising avenues for future research.

2. Methodology

Recently in the natural language processing field Bayesian models have been proposed to tackle a variety of problems, and have been found to be particularly effective in word segmentation [12, 13]. The model we use in this paper is a unigram Dirichlet process model. Using this approach to perform bilingual segmentation for the general case of machine translation with re-ordering would be a challenging undertaking, however for transliteration where the sequence lengths are short and under the assumption that there is no re-ordering, it is feasible to tackle the bilingual segmentation problem directly without the need for specialized optimization or annealing (we do use a block sampling algorithm, and a dynamic programming algorithm).

2.1. Joint Source-channel Model

Let us assume we are given a bilingual corpus consisting of a source sequence $\bar{\mathbf{s}}_1^M = <s_1, s_2, \ldots, s_M>$ and a target sequence $\bar{\mathbf{t}}_1^N = <t_1, t_2, \ldots, t_N>$. We distinguish sequences of characters from single characters by using a boldface font with an overbar.

We adopt the joint source-channel model of [6] as the underlying generative model, and we make the additional assumption that the segments are independent of each other (our approach can easily be extended to model these dependencies at the expense of some additional complexity, see [13]). Under this model, the corpus is generated through the concatenation of *bilingual sequence-pairs* (we will use this term repeated throughout this paper to refer to corresponding sequences of source and target graphemes, as defined below).

A bilingual sequence-pair is a tuple $(\bar{\mathbf{s}}, \bar{\mathbf{t}})$ consisting of a sequence of source graphemes together with a sequence of target graphemes $(\bar{\mathbf{s}}, \bar{\mathbf{t}}) = (<s_1, s_2, \ldots s_i>, <t_1, t_2, \ldots, t_j>)$.

The corpus probability is simply the probability of all possible derivations of the corpus given the set of bilingual sequence-pairs and their probabilities.

$$\begin{aligned} p(\bar{\mathbf{s}}_1^M, \bar{\mathbf{t}}_1^N) &= p(s_1, s_2, \ldots, s_M, t_1, t_2, \ldots, t_N) \\ &= \sum_{\gamma \in \Gamma} p(\gamma) \end{aligned} \tag{1}$$

where $\gamma = ((\bar{\mathbf{s}}_1, \bar{\mathbf{t}}_1), \ldots, (\bar{\mathbf{s}}_k, \bar{\mathbf{t}}_k), \ldots, (\bar{\mathbf{s}}_K, \bar{\mathbf{t}}_K))$ is a derivation of the corpus characterized by its co-segmentation, and Γ is the set of all derivations (co-segmentations) of the corpus.

The probability of a single derivation is given by the product of its component bilingual sequence-pairs:

$$p(\gamma) = \prod_{k=1}^{K} p((\bar{\mathbf{s}}_k, \bar{\mathbf{t}}_k)) \tag{2}$$

The corpus for our experiments is segmented into bilingual word-pairs. We therefore constrain our model such that both source and target character sequences of each bilingual sequence-pair in the derivation of the corpus are not allowed to cross a word segmentation boundary. Equation 2 can therefore be arranged as a product of word-pair w derivations of the sequence of all word-pairs $\mathcal{W}$ in the corpus.

$$p(\gamma) = \prod_{w \in \mathcal{W}} \prod_{(\bar{\mathbf{s}}_k, \bar{\mathbf{t}}_k) \in \gamma_w} p((\bar{\mathbf{s}}_k, \bar{\mathbf{t}}_k)) \tag{3}$$

where γ_w is a derivation of bilingual word-pair w.

2.2. Unigram Dirichlet Process Model

A Dirichlet process is a stochastic process defined over a set S (in our case, the set of all possible bilingual sequence-pairs) whose sample path is a probability distribution on S.

The Dirichlet process model we use in our approach is a simple model that resembles the cache models used in language modeling [14]. Intuitively, the model has two basic components: a model for generating an outcome that has already been generated at least once before, and a second model that assigns a probability to an outcome that has not yet been produced. Ideally, to encourage the re-use of model parameters, the probability of generating a novel bilingual sequence-pair should be considerably lower then the probability of generating a previously observed sequence pair. This is a characteristic of the Dirichlet process model we use and furthermore, the model has a preference to generate new

sequence-pairs early on in the process, but is much less likely to do so later on. In this way, as the cache becomes more and more reliable and complete, so the model prefers to use it rather than generate novel sequence-pairs. The probability distribution over these bilingual sequence-pairs (including an infinite number of unseen pairs) can be learned directly from unlabeled data by Bayesian inference of the hidden co-segmentation of the corpus. The ability of the model to assign a probability to any unseen sequence-pair gives the technique the ability to score candidate training data.

The underlying stochastic process for the generation of a corpus composed of bilingual phrase pairs γ is usually written in the following from:

$$\begin{aligned} G|_{\alpha,G_0} &\sim DP(\alpha, G_0) \\ (\bar{\mathbf{s}}_k, \bar{\mathbf{t}}_k)|G &\sim G \end{aligned} \tag{4}$$

G is a discrete probability distribution over the all bilingual sequence-pairs according to a *Dirichlet process prior* with *base measure* G_0 and concentration parameter α. The concentration parameter $\alpha > 0$ controls the variance of G; intuitively, the larger α is, the more similar G_0 will be to G.

2.2.1. The Chinese Restaurant Process

Unfortunately it is not possible to estimate G directly, since there are an infinite number of possible bilingual sequence-pairs, so instead we integrate over its possible values. To do this we cast the bilingual sequence-pair generation process as an instance of the Chinese Restaurant Process (CRP) [15]. According to this representation, every bilingual sequence-pair corresponds to the dish served at its table in a potentially infinite set of tables in a Chinese restaurant. The number of customers seated at each table represents the cumulative count of the bilingual sequence-pair. A new customer to the restaurant can take a seat at an occupied table with a probability proportional to the number of customers at that table, and must eat that table's dish, or can take a seat at an unoccupied table with a probability proportional to a constant, in which case they must eat a dish (a bilingual sequence-pair) chosen by the chef (in this analogy the chef's choice is in accordance with the base distribution G_0).

2.2.2. The Base Measure

For the *base measure* that controls the generation of novel sequence-pairs, we use a joint spelling model that assigns probability to new sequence-pairs according to the following joint distribution:

$$\begin{aligned} G_0((\bar{\mathbf{s}}, \bar{\mathbf{t}})) &= p(|\bar{\mathbf{s}}|)p(\bar{\mathbf{s}}||\bar{\mathbf{s}}|) \times p(|\bar{\mathbf{t}}|)p(\bar{\mathbf{t}}||\bar{\mathbf{t}}|) \\ &= \frac{\lambda_s^{|\bar{\mathbf{s}}|}}{|\bar{\mathbf{s}}|!} e^{-\lambda_s} v_s^{-|\bar{\mathbf{s}}|} \times \frac{\lambda_t^{|\bar{\mathbf{t}}|}}{|\bar{\mathbf{t}}|!} e^{-\lambda_t} v_t^{-|\bar{\mathbf{t}}|} \end{aligned} \tag{5}$$

where $|\bar{\mathbf{s}}|$ and $|\bar{\mathbf{t}}|$ are the length in characters of the source and target sides of the bilingual sequence-pair; v_s and v_t are that vocabulary (alphabet) sizes of the source and target languages respectively; and λ_s and λ_t are the expected lengths of source and target.

According to this model, source and target sequences are generated independently: in each case the sequence length is chosen from a Poisson distribution, and then the sequence itself is generated given the length. Note that this model is able to assign a probability to arbitrary bilingual sequence-pairs of any length in source and target sequence, but favors shorter sequences in both.

Following [12] we assign the parameters λ_s, λ_t and α, the values 2, 2 and 0.3 respectively. Ideally these parameters should be learned from the data, however in our experiments the settings were sufficient to give a useful co-segmentation of the training corpus. Moreover, the system proved to be insensitive to changes in these parameters in a set of pilot experiments, converging to very similar final iteration samples for a range of parameter settings.

2.2.3. The Generative Model

The generative model is given in Equation 6 below. The equation assignes a probability to the k^{th} bilingual sequence-pair $(\bar{\mathbf{s}}_k, \bar{\mathbf{t}}_k)$ in a derivation of the corpus, given all of the other sequence-pairs in the history so far $(\bar{\mathbf{s}}_{-k}, \bar{\mathbf{t}}_{-k})$. Here $-k$ is read as: "up to but not including k".

$$\begin{aligned} p((\bar{\mathbf{s}}_k, \bar{\mathbf{t}}_k))|(\bar{\mathbf{s}}_{-k}, \bar{\mathbf{t}}_{-k})) = \\ \frac{N((\bar{\mathbf{s}}_k, \bar{\mathbf{t}}_k)) + \alpha G_0((\bar{\mathbf{s}}_k, \bar{\mathbf{t}}_k))}{N + \alpha} \end{aligned} \tag{6}$$

In this equation, N is the total number of bilingual sequence-pairs generated so far (the number of customers so far), $N((\bar{\mathbf{s}}_k, \bar{\mathbf{t}}_k))$ is the number of times the sequence-pair $(\bar{\mathbf{s}}_k, \bar{\mathbf{t}}_k)$ has occurred in the history (the number of people seated at its table). G_0 and α are the base measure and concentration parameter as before.

3. Bayesian Inference

3.1. Gibbs Sampling

We used a blocked version of a Gibbs sampler for training. In [14] they report issues with mixing in the sampler that were overcome using annealing. In [13] this issue was overcome by using a blocked sampler together with a dynamic programming approach. Our algorithm is similar to that of [13], and we found our sampler converged rapidly without annealing (see Figure 2). The number of iterations was set by hand after observing the convergence behavior of the algorithm in pilot experiments. We used a value of 30 iterations through the corpus in all our experiments.

The sampling algorithm is shown in Algorithm 1 and the iterative component proceeds as follows. Firstly the training set of bilingual word-pairs is permuted randomly, and a bilingual word-pair is sampled from this permutation without replacement. Secondly, a probability distribution over

Input: Random initial corpus segmentation
Output: Unsupervised co-segmentation of the corpus according to the model
foreach *iter=1 to NumIterations* **do**
  foreach *bilingual word-pair* $w \in$ *randperm*($\mathcal{W}$) **do**
    foreach *co-segmentation* γ_i *of* w **do**
      Compute probability $p(\gamma_i|h)$
      where h is the set of data (excluding w) and its hidden co-segmentation
    end
    Sample a co-segmentation γ_i from the distribution $p(\gamma_i|h)$
    Update counts
  end
end

Algorithm 1: The blocked Gibbs sampling algorithm.

all possible co-segmentations of the chosen bilingual word-pair is calculated by obtaining probabilities with respect to a model that does not include the bilingual word-pair, its previous segmentation information and respective counts. Due to the short sequence lengths involved in transliteration, it is possible to use a brute force approach to calculate this distribution, however for efficiency we extended the forward filtering/backward sampling (FFBS) dynamic programming algorithm of [13] to deal with bilingual segmentation. We implemented this algorithm graphically as explained below.

We use a segmentation graph (shown in Figure 1) to guide the process. This directed graph is a compact representation of all possible ways in which to co-segment a bi-lingual pair. Each node represents a set of partial co-segmentation hypotheses of the whole sequence that share the same sequences of source and target tokens, and each arc represents the bilingual phrase pair used to transition from the tail of the arc to the head. In the figure the arcs are labelled with the log-probability of this sequence-pair (given by the model in Equation 6), therefore the log-probability of a full segmentation hypothesis is given by the sum of the arc labels on the respective path from the source node "<s>" to the sink node "abba アッバ". The most probable co-segmentation is indicated with bold arcs in the figure and corresponds to the segmentation a/ア b/ッ ba/バ, this is reasonable since both "a" and "ba" are associated with their phonetic equivalents in Japanese, and the Japanese "ッ" indicates that the consonant immediately to the right is to be repeated. The least probable segmentation in the graph is given by abb/ア a/ッバ. The log-probabilities in the graph are real values taken from the third iteration of the training, and here the most probable segmentation is already by far the most likely.

Nodes in the graph can have multiple in- and out-degree. Two nodes are combined when the unsegmented part of the bi-lingual sequence pair is the same for both, giving rise to a compact, efficient representation.

The FFBS algorithm operates directly on the segmentation graph, and has two steps. The *forward filtering* step, calculates for each node in the graph, the probability of the subgraph (including the node itself) to the left of the node, back to the source node. This probability α, is stored in the node itself (these α's are shown in Figure 1). This process proceeds recursively in a depth-first post-order traversal of the graph, starting at the sink node. Nodes for which the probability has been calculated are marked as done, ensuring α gets calculated only once for the node.

The *backward sampling* step samples a derivation of the bi-lingual word pair according to the probability distribution over all possible segmentations. This is done easily using the α values stored in the graph by the forward filtering process. The backward sampling also proceeds recursively from the sink node. For each incoming arc, the probability of including that arc in the sample is given by the product of the arc probability and the α value at the tail of the arc. This value is calculated for each incoming arc, and one arc from the set is sampled according to the probability distribution over the arcs. The sampling procedure is called recursively on the tail of the sampled arc until the source node of the graph is reached. The sequence of arcs traversed defines the sampled derivation of the bi-lingual pair for the current iteration of the training process, and this sample is in accordance with the probability distribution over all derivations with respect to the model.

3.2. Sequence-pair Extraction

During the phrase-table generation process of a typical phrase-based SMT system, GIZA++ is run twice to generate alignments at the word level, from source-to-target and from target-to-source. Following this step, the $grow\text{-}diag\text{-}final\text{-}and$ procedure is used to extract *all phrases* consistent with the word alignments arising from the two GIZA++ runs. When building a phrase-table from the alignment achieved at final iteration of our Gibbs sampling procedure, we use a much simpler heuristic that is in the same spirit to derive a larger set of phrases consistent with the initial co-segmentation. Our experiments show that this is a necessary step that considerably improves system performance.

The algorithm we use for phrasal extraction from the co-segmented corpus is as follows: within a single bilingual word-pair, agglomerate all contiguous bilingual sequence-pairs in all possible ways, but limit the size of the agglomer-

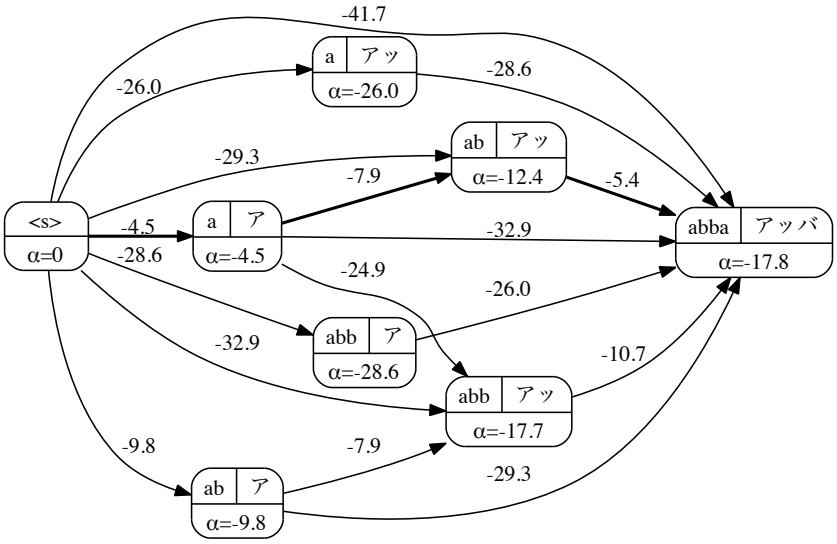

Figure 1: A graph representing all possible co-segmentations of the character sequences "abba" in English and "アッバ" in Japanese. The α labels on the nodes represent the log-probability of subgraph (including the node itself) to the left of the node. The labels on the arcs are the log-probabilities of bi-lingual phrase pairs used to transition from tail-to-head, and are given by the model of Equation 6.

ated source and target phrases to match the *maximum phrase length* parameter used to train the SMT system (this was set to 7 in our experiments). This is not strictly necessary, but we performed this step to keep the phrase-table generated from our Bayesian segmentation comparable to that generated by the baseline system.

4. Experiments

4.1. Baseline System

For our experiments we use the phrase-based machine translation techniques introduced by [5], integrating our models within a log-linear framework [16]. Word alignment was performed using GIZA++ [4] and sequence-pair extraction using the MOSES [5] tools. The decoder used was an in-house phrase-based machine translation decoder that operates according to the same principles as the publicly available MOSES [5] SMT decoder.

In these experiments 5-gram language models built with Witten-Bell smoothing were used. The system was trained in a standard manner, using a minimum error-rate training (MERT) procedure [17] with respect to the BLEU score on the held-out development data to optimize the log-linear model weights.

Rama and Gali [18], evaluated several techniques for sequence-pair extraction for transliteration and found the $grow\text{-}diag\text{-}final\text{-}and$ heuristic to be the most effective, we therefore adopt this method in the baseline system our experiments.

4.1.1. Decoding Constraints

The experiments reported in this paper were conducted using a beam width of 100, with no stack thresholding, and a strictly monotone decoding process.

4.2. Experimental Data

Our training data consisted of 27993 bilingual single word-pairs that were used in the NEWS2010 workshop transliteration shared task. The development data consisted of 3606 bilingual word-pairs drawn from the same sample. The evaluation data consisted of a further 1935 bilingual word-pairs not contained in the other two data sets. The corpus statistics for the three corpora are given in Table 1.

Corpus	word-pairs	Characters	
		En	Ja
Training	27993	188941	131275
Development	3606	24066	16651
Evaluation	1935	11863	8199

Table 1: Statistics of the English-Japanese bilingual corpora.

We used the data to train a phrase-based SMT system to perform transliteration from English to Japanese. We trained our Dirichlet process model on the same parallel data

set, and extracted transliteration phrase-tables from the co-segmentation of the corpus at the final iteration (iteration 30).

4.3. Training Procedure

For the Gibbs sampling, we chose to start the sampling from a random co-segmentation of the corpus. That is, for each bilingual word-pair in the corpus, a single co-segmentation was sampled from a uniform distribution over all possible co-segmentations of the pair. We believe that it might be advantageous, and certainly more efficient to start the sampling from a more intelligent starting point, for example one derived from a pre-processing pass of GIZA++. However, the training was able to arrive at a good segmentation (by visual inspection) of the training corpus, its usefulness being borne out by the experimental results in the next section.

4.4. Evaluation Procedure

The results presented in this paper are given in terms of official evaluation metrics used in the NEWS2010 transliteration generation shared task [19]. In our results, ACC refers to the top-1 accuracy score, that measures the percentage of the time the top hypothesis from the system exactly matches the reference. F-score measures the distance of the best hypothesis from the reference transliteration; the reader is referred to the workshop white-paper [19] for more details. For brevity, we only report our results in terms of ACC and F-score in this paper, but the results in terms of the other NEWS2010 metrics have the same character.

5. Results

5.1. Training

The convergence of the algorithm during the training procedure is shown in Figure 2 which plots the log-probability of the sampled derivation at the end of each pass through the training corpus against iteration. It can be seen from the graph that the system rapidly improves from the poor initial segmentation, and thereafter continues to gradually improve. The log-probability of the initial random co-segmentation was -1.5e06 and is omitted.

Comparing the segmentations of at various iterations gives some insight here. At iteration 29, 90% of all the words in the corpus are bilingually segmented in an identical manner to those at iteration 30. Since this is a sampling process, the 10% that differ may be explained by different choices made in the sampling process. At iteration 3, 87.5% of word-pairs are already segmented in an identical manner to iteration 30.

5.2. Evaluation

Our results on the English-to-Japanese transliteration task are summarized in Table 2. It is clear from the table that using sequence-pairs from only the sample at the final iteration of the training produces gave lower performance than the baseline system. The phrase-table derived in this way contained

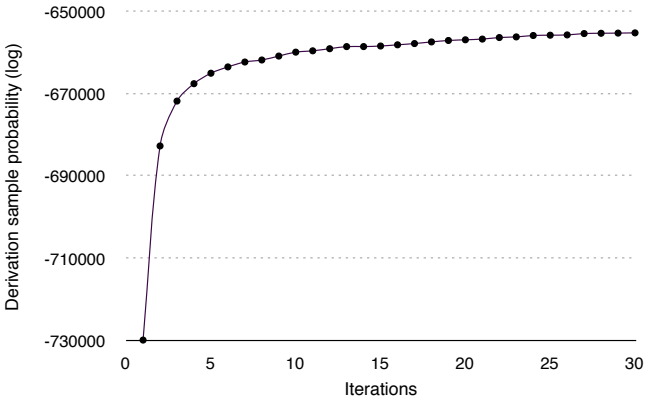

Figure 2: The evolution of the log-probability of the sampled derivation with respect to the training iteration.

only 3372 sequence-pairs as opposed to over 140,000 in the phrase-table extracted from the GIZA++ alignments. Moreover these sequence-pairs were very short compared to those in from the baseline system's phrase-table: approximately 3 characters in both source and target on average, compared to around 5 characters for the baseline system.

When a phrase-table built from agglomerations of the same set of sequence-pairs was used, a much larger phrase-table of around 100,000 phrases resulted, with sequence-pairs that are comparable in size to those of the baseline, around 5 characters. On the transliteration task, this phrase-table gave an improvement of approximately 1% in ACC over the baseline system, from a phrase-table that was about 30% smaller in size. Moreover, since the sequence-pairs are concatenations of 3372 component sequence-pairs, this model could be stored very compactly if necessary. Further gains were obtained by interpolating the agglomerated model together with the baseline model. We believe this gain may be due to the effect of smoothing.

Our experiments were designed to favor the baseline model since the system was tuned using the MERT procedure with its own phrase-table. It is possible that our proposed system would have obtained a higher score if tuned with its own phrase-table, however we chose not to as this would have introduced additional variance from the differences in the two MERT search processes into the results. In a second experiment we collected counts for the sequence-pairs over multiple iterations of the training process: from iteration 5 (the *burn-in*) to the final iteration. This resulted in a 37% larger phrase-table, but surprisingly did not realize any notable improvement in performance.

It is interesting to note that the system's performance was improved dramatically simply by grouping the phrases into larger units. This highlights one of the advantages of the phrase-based translation approach. The agglomerated model, because of the way it was constructed, is not able to generate anything the simpler model cannot, but when larger sequence-pairs are used to build the target sequence the char-

Phrase Extraction Model	ACC	F-score	Phrase-table Entries	Avg. Phrase Length En	Ja
GIZA++ and *grow-diag-final-and*	0.313	0.745	143382	5.41	4.80
Bayesian Segmentation	0.278	0.726	3372	2.60	2.75
Bayesian Segmentation (+agglomerated)	0.323	0.748	102507	5.54	4.83
Bayesian Segmentation (+integrated)	0.329	0.752	164258	5.46	4.81

Table 2: The experimental results for the three systems together with some statistics of their phrase-tables. Here *+agglomerated* means the sequence-pairs were extracted by agglomeration from a single sample at the end of the training. In *+integrated* the phrase-tables from the baseline system and the agglomerated system were linearly interpolated with equal weights. Differences between systems were all found to be significant by paired t-testing at a level of 0.05, except for the ACC scores for the agglomerated and integrated systems.

acters in the phrase carry with them the *implicit context* of the other characters in the phrase, all of which have occurred together in the same context in the training corpus. In the model with the unagglomerated sequence-pairs, this role is performed mainly by the language model. In spite of the fact that we used a 5-gram language model the system clearly benefited from a model that contained longer sequence-pairs as the basic translation unit.

5.3. Decoding Consistency

We ran an experiment to investigate the reasons for the improvements in system performance. Our hypothesis was that the Bayesian system had produced a phrase table that led to a more consistent decoding process. This was based on the belief that the fact that the Dirichlet process model strongly encourages reuse of the bilingual sequence-pairs it discovers. This should result in a more compact phrase-table, and should entail that similar words in the corpus are likely to be decoded in more homogenous fashion. To test the hypothesis we modified the machine translation decoder to count the number of *types* of bilingual sequence-pair used to decode the evaluation data, and re-ran the English-Japanese transliteration experiment that showed the largest gain in performance. We found that the decoding process that used the phrase-table generated from our Bayesian model (with agglomerated sequence-pairs) used a total of 3496 unique sequence-pairs, whereas decoding using the phrase-table extracted using GIZA++ and *grow-diag-final-and* required a total of 3970 phrase pairs during the decoding process, supporting our hypothesis. The 3496 sequence-pairs from the Bayesian model's phrase-table, could be further analysed into 1289 component bilingual pairs that were present in the segmentation in the sample taken at the end of the training process.

6. Conclusion

In this paper we have presented a novel Bayesian bilingual co-segmentation scheme and applied it to the task of phrase-table generation for transliteration by phrase-based statistical machine translation. Traditional models of phrasal alignment rely on maximum likelihood training coupled with the EM algorithm, but have serious issues with overfitting the training data. Because of these issues, alignment is typically performed in a one-to-many manner from source-to-target and from target-to-source and the phrase extraction process proceeds heuristically from an alignment table. Our approach offers the ability to align the training data in a many-to-many fashion directly using Bayesian techniques that offer a simple yet elegant solution to the issues inherent in maximum likelihood training. In addition, our approach is symmetrical with respect to source and target, and also with respect to the word order of the corpus.

We investigated the quality of the bilingual phrasal alignment achievable with unsupervised Bayesian co-segmentation, and designed experiments to compare directly to a standard GIZA++/*grow-diag-final-and* phrase extraction procedure by constructing a phrase-table from samples arising from the Gibbs sampling training procedure. Our experiments show that the Bayesian approach is able to produce a smaller phrase-table that can offer comparable or higher transliteration performance than the baseline system.

Furthermore, our technique offers other benefits: one example being that it provides a full co-segmentation of the training corpus at the end of training which can be used to directly train a joint sequence model. This contextual information is a key feature in joint sequence transliteration models such as [6], and is currently missing from the phrase-based SMT-based transliteration systems. Another virtue of our approach stems from the fact that the Dirichlet process model is able to assign a probability to any bilingual word pair. We believe this type of model in has considerable potential utility in transliteration mining and corpus filtering, since it provides a principled way of scoring any potential transliteration candidate.

In future research we would like to investigate the effect of introducing a joint sequence model feature into a phrase-based SMT-based transliteration system. We also plan to improve the underlying Dirichlet process model in order to better model the data, moving to higher-order and hierarchical models.

7. References

[1] A. Finch and E. Sumita, "Phrase-based machine transliteration," vol. 1, Hyderabad, India, 2008.

[2] T. Rama and K. Gali, "Modeling machine transliteration as a phrase based statistical machine translation problem," in *NEWS '09: Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration*. Morristown, NJ, USA: Association for Computational Linguistics, 2009, pp. 124–127.

[3] S. Noeman, "Language independent transliteration system using phrase based smt approach on substrings," in *NEWS '09: Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration*. Morristown, NJ, USA: Association for Computational Linguistics, 2009, pp. 112–115.

[4] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[5] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowa, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: open source toolkit for statistical machine translation," in *ACL 2007: proceedings of demo and poster sessions*, Prague, Czeck Republic, June 2007, pp. 177–180.

[6] H. Li, M. Zhang, and J. Su, "A joint source-channel model for machine transliteration," in *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, 2004, p. 159.

[7] D. Yang, P. Dixon, Y.-C. Pan, T. Oonishi, M. Nakamura, and S. Furui, "Combining a two-step conditional random field model and a joint source channel model for machine transliteration," in *NEWS '09: Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration*. Morristown, NJ, USA: Association for Computational Linguistics, 2009, pp. 72–75.

[8] D. Marcu and W. Wong, "A phrase-based, joint probability model for statistical machine translation," in *In Proceedings of EMNLP*, 2002, pp. 133–139.

[9] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Commun.*, vol. 50, no. 5, pp. 434–451, 2008.

[10] P. Blunsom, T. Cohn, C. Dyer, and M. Osborne, "A gibbs sampler for phrasal synchronous grammar induction," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL*. Suntec, Singapore: Association for Computational Linguistics, August 2009, pp. 782–790.

[11] J. Wuebker, A. Mauser, and H. Ney, "Training phrase translation models with leaving-one-out," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden: Association for Computational Linguistics, July 2010, pp. 475–484.

[12] J. Xu, J. Gao, K. Toutanova, and H. Ney, "Bayesian semi-supervised chinese word segmentation for statistical machine translation," in *COLING '08: Proceedings of the 22nd International Conference on Computational Linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, 2008, pp. 1017–1024.

[13] D. Mochihashi, T. Yamada, and N. Ueda, "Bayesian unsupervised word segmentation with nested pitman-yor language modeling," in *ACL-IJCNLP '09: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1*. Morristown, NJ, USA: Association for Computational Linguistics, 2009, pp. 100–108.

[14] S. Goldwater, T. L. Griffiths, and M. Johnson, "Contextual dependencies in unsupervised word segmentation," in *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, 2006, pp. 673–680.

[15] D. J. Aldous, "Exchangeability and related topics," in *École d'été de probabilités de Saint-Flour, XIII—1983*, ser. Lecture Notes in Math. Berlin: Springer, 1985, vol. 1117, pp. 1–198.

[16] F. J. Och and H. Ney, "Discriminative training and maximum entropy models for statistical machine translation," in *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, 2002, pp. 295–302.

[17] F. J. Och, "Minimum error rate training for statistical machine trainslation," in *Proceedings of the ACL*, 2003.

[18] T. Rama and K. Gali, "Modeling machine transliteration as a phrase based statistical machine translation problem," in *In Proc. ACL/IJCNLP Named Entities Workshop Shared Task*, 2009.

[19] M. Z. Haizhou Li, A Kumaran and V. Pervouchine, "Whitepaper of news 2010 shared task on transliteration generation," in *In Proc. ACL Named Entities Workshop Shared Task*, 2010.